Summary of doctoral dissertation

NETWORK COMMUNICATION. BIG DATA INFORMATION SOURCES

KOMUNIKACJA SIECIOWA. ŹRÓDŁA INFORMACJI BIG DATA

Author: Dariusz Jaruga

Warszawa, październik 2019

STRESZCZENIE

Komunikacja sieciowa, fundamentalna składowa nauk o mediach, niezależnie od jej kategorii: bezpośrednia-synchroniczna, pośrednia-asynchroniczna, aktywna, autonomicznych maszyn, generuje ogromne ilości danych cyfrowych określanych jako Big Data. Dane te, udostępnione w sieci za pośrednictwem różnych usług i formatów, stanowią cenny surowiec badawczy, z którego możliwe jest pozyskanie nowych użytecznych informacji. Przedmiotem pracy jest problematyka źródeł dużych zasobów informacyjnych Big Data, ich zautomatyzowanego kolekcjonowania i przetwarzania do postaci umożliwiającej ilościową analizę. Przedstawiono cztery różne badania, ilustrujące potencjał drzemiący w informacyjnych zasobach Big Data, zrealizowane za pomocą narzędzia teleinformatycznego robota Big Data Jazon, skonstruowanego w ramach niniejszej dysertacji.

ABSTRACT

A fundamental component of media science is network communication. All forms of network communication – direct and synchronous, intermediate and asynchronous, and active and autonomous – generate huge amounts of digital data called Big Data. This data is available on the network through various services and formats and is very valuable in terms of research, as it can be used to acquire new useful information. This work discusses the sources of Big Data and their automated collection and processing that allow analyses to be made. Four different studies that illustrate the potential of Big Data were conducted as part of this dissertation. These were implemented using a new ICT tool – the Big Data robot, Jazon.

SŁOWA KLUCZOWE

Big Data, rafinacja informacji, komunikacja sieciowa, źródła informacji, robot, sentymenty, kolekcjonowanie danych,

KEY WORDS

Big Data, data refining, network communication, sources of information, robot, sentiments, data collection.

INDEX

Justification for topic

Information is an essential component of media science. It is of particular importance to internet resources and breakthrough technologies that collect, exchange, and process information. As a result, the large-scale information space is becoming an area of new research.

Through the use of machines, it has become possible to transform large amounts of data into high quality, new information. Machines are the foundation of all forms of network communication, which is a medium of information exchange between humans, computers, and both humans and computers[1]. Given the communicative nature of the network, it seems appropriate to highlight the sources of information found within network communication and the content they contain[2]. Once found, the content extracted from these sources can be normalized, collected, and transferred to the recipient – a machine (for further processing) or a human (for direct use) [3]. The work focuses on the issue of collecting information and transferring it for further processing known as refining information.

Common access to relatively cheap as well as efficient and reliable ICT tools has resulted in the modern world and the processes in it to be described, recorded, and filmed. The amount of multimedia content produced and published by ordinary people – amateurs – and professionals has reached dimensions that are hard to even imagine. Contractual network users and the machines and devices associated with them today send much more information than humans. The subject of this work is the issue of how to select information in digital form and transfer it, after further processing, to recipients.

Users' involvement in the network includes a wide range of activities that produce information. These activities include writing a blog, commenting, participating in discussions, assessing other creators, expressing opinions, uploading their own musical works or videos, and communicating with established contacts. Data is also generated by devices without any direct human intervention. This includes data from video recorders and vehicle fleet and navigation systems that automatically save the position of the vehicle, data generated through direct communication between machines (Internet of Things, IoT, Machine-to-Machine, M2M), or data from Near Field Communication (NFC, which is currently used in proximity cards and smartphones to exchange digital

[1] Włodzimierz Gogołek, *Komunikacja sieciowa: uwarunkowania, kategorie i paradoksy* (Warszawa: Oficyna Wydawnicza ASPRA JR., 2010).

[2] The concept of 'network' in this work should be understood as the internet and its services.

[3] Normalization consists of transforming various forms of digital information downloaded from the network into a unified internal data format, allowing information to be stored in the database for that particular source of information.

data between devices up to 20 cm apart[4]). All this information is raw material that can be used to extract new data, and thus, can be used by computers to independently produce messages addressed to people and other machines.

The emergence of huge information resources and the development of communication technologies cause changes in social and economic areas. In the social area, we talk about the information society, and in the field of economics, about a knowledge-based economy. The combination of knowledge and information is one of the factors that shape economic and social development. Knowledge obtained from the network contributes to the creation of new, useful products and services. Thus, knowledge and information is valuable capital for the development of the modern world. The increase in amount of information enriches knowledge about current problems and important phenomena occurring in society. The catalyst for the increase in this knowledge is the ability to analyze large information resources known as Big Data.

Modern technology has enabled two-way communication between its users, and social networks have enabled large groups of people to efficiently communicate. The content of the collected messages have become a valuable source of information that can be further analyzed and refined. Information from these messages allows one to draw conclusions about every sphere of modern life (e.g. culture, economy, and politics). By analyzing collected data, changes in the field of economics, social and political phenomena, crises, revolutions, epidemics, etc., can be anticipated.

Network resources (Big Data) are useful in scientific research conducted by not only research centers and universities, but also, and more importantly, enterprises, institutions, and public administrative bodies. Professional data analysis can be very useful for people and, if used properly, translates into improved quality of life. There is also the other side of the coin, where results of Big Data analysis can be used against people.

The fundamental task of refining information is not to collect large amounts of data, but rather to choose which of many sources of information are useful for further analysis. For this purpose, an appropriate tool is necessary. This forms an important part of this dissertation. The results of analysis of data obtained from the network can be used in various areas, including business management, economic or epidemiological assessments, and detecting adverse social phenomena or symptoms of a crisis. It can also be used to see approaching changes in the modern world with a certain probability. Big Data analysis can therefore have a direct impact on the way public utilities

[4]W. Pawłowicz, *Perspektywy komunikacji M2M*
http://www.computerworld.pl/news/380926/Perspektywy.komunikacji.M2M.html
[dostęp 26-07-2016], oraz About Near Field Communication.

operate, e.g. stations, airports, offices, or commercial companies. It can also be used to find the right solution for potential investments, e.g. road, rail connections, or appropriate preventive healthcare.

The research carried out as part of this work - in addition to verifying the functionality of the robot - has a consequence: bringing out information hidden in large resources[5]. Information gained in this way can allow people to achieve practical goals, including examining the image of a person or company, developing recommendations for existing or new products and services, innovation in B+R+I (Build + Research + Innovation), finding expected or desired product characteristics, or attempting to determine the probability of specific social phenomena in the future. This new information, or new links in the communication process, is obtained from the information collected by the Big Data robot, Jazon. This information is related directly or indirectly to the topic chosen by the researcher. The main sources of data for social research are social networking sites, forums, blogs, and relevant internet services. For the area of innovation, valuable sources of primary information are scientific and industry publications, and descriptions of patents and tender procurement items.

The information extracted by the Big Data robot provides the link between Big Data and the recipient. The Big Data robot is the first link of the information transmission, as it collects and refines the information, allowing research to be conducted. Four studies were used to analyze the information collected by the Big Data robot. Each study concerned a different topic and involved either structured or unstructured data, and current or historical data[6]. The studies were very different, which proves that the robot is universal, and illustrate the robot's many applications. The robot has been used repeatedly since May 2015 for various studies in the fields of social sciences, economics, innovation, and science and technology. The results of the studies show that the Big Data robot can be successfully used to refine information, which can then be used to describe and analyze society and science, and any changes taking place within them.

[5] The Big Data robot, Jazon is a tool that was built as a product of the dissertation.
[6] Structured data is information with a defined structure, e.g. a table in a database. Unstructured data is data without a defined structure, e.g. the text content of articles.

Placing the topic of the work in the sciences

The subject of this dissertation is in the field of social sciences, in particular in the new independent discipline of media science, whose existence was confirmed by the Regulation of the Minister of Science and Higher Education of August 8, 2011, on areas of knowledge, fields of science, and art, as well as scientific and artistic disciplines[7]. According to prof. Jabłonowski and dr. hab. Gackowski, media science concerns three areas of science - social, humanities, and technical[8]. Considering that the dynamic development of network applications and services enable multifaceted communication in all areas of science, the role of media sciences is important in each of them.

As part of this dissertation, an ICT tool has been developed and manufactured. Due to the functionality of this tool, this work also falls under the discipline of computer science. Without computer science, it would not be possible to study large data sets known as Big Data. Thus, in almost all areas of science, the importance of technical sciences and technological aspects in media transformation is growing. New media and the internet have expanded existing forms of communication.

This work is multi-faceted and attempts to show valuable sources of information obtained from network resources. It concerns, inter alia, the impact of new technologies on creators and recipients of network information and the potential of data created by internet users. This work combines several disciplines such as media studies, computer science, bibliology, politics, cognition and social communication, pedagogy, psychology, sociology, economics, finance, management, and law. Depending on the type of data collected, various aspects of phenomena occurring in the modern world can be studied, affecting individual people as well as entire groups. The research material collected by the robot, subjected to subsequent analysis, allows one to conclude and predict the near future and gives the opportunity to better understand the changes taking place in the world around us – the world of people.

[7] *Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z dnia 8 sierpnia 2011 r. w sprawie obszarów wiedzy, dziedzin nauki i sztuki oraz dyscyplin naukowych i artystycznych* [na:] http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20111791065, udostępniono 20 maja 2018 r.

Aim of the work, research problem, hypothesis

The enormous progress in the field of ICT, which has created a field for various forms of communication, prompts us to explore the not fully recognized area of social phenomena occurring in the network. There are two purposes of this dissertation. The first is to develop a methodology for the construction of an ICT tool – a Big Data robot – that consists of hardware and dedicated software for collecting and analyzing network resources known as Big Data. The second is to create a dedicated Big Data database in which information can be refined and used for further exploration of large information resources.

The main hypothesis of the dissertation is that the developed methodology together with the ICT tool is a universal, parameterized system that enables new information to be effectively obtained from large information resources (Big Data).

Research Method

As the subject of this dissertation is a methodology for collecting information, ready-made solutions on the market were analyzed. These solutions included available services as well as tools that could be used to build the methodology and finally build a robot based on it. Software tools included program libraries, e.g. RSS feeds, ready systems such as information brokers, or client applications for database support. The quality and usefulness of the different solutions were assessed based on information contained in the network, such as analytical reports, manufacturer's documentation, case studies of their application, and opinions of end users and people using a given tool to build their own systems. The system's maturity, including its history, popularity in rankings, stability, security, and total maintenance costs were also taken into account. The functionality of individual solutions was also compared. For example, for the information broker, the following parameters were analyzed: documentation, possibility of cooperation with other solutions, interface intuitiveness, user community, development history, entities using the solution, and license. The laboratory environment was prepared before the final selection. In this, a pre-selected solution was tested, which was then selected or rejected, leading to the next one being analyzed. It was found that the automatic information-collection products and software tools currently available on the market could be used to build the robot, and the Big Data robot, Jazon, was constructed.

Big Data research is closely related to technology, even in research in the field of social sciences. Technology must be used in order to analyze Big Data. Studies performed on large data sets often require information to be refined. Therefore, in research carried out with the Big Data robot, data was collected and then refined.

Refining information consists of six stages: formulating the goal, developing a set of names and sentiments, collecting entries from pages, verifying sentiments, calculating attendance, and interpreting results[9]. Each of these stages is described in detail in the dissertation.

Rating sources

This dissertation was prepared on the basis of literature studies and empirical research, including experiments. Literature in the area of social sciences, including media science and information science was used. Of particular importance was literature relating to new communication technologies. To assist with the construction of the Big Data robot, literature from the fields of computer science, mathematics, and cybernetics was used. This included technical documentation for individual elements included in the Big Data robot, and books and scientific articles in Polish and English. In addition, legal acts, statistical data, reports, rankings, specialized websites, specialized forums and internet blogs were also used. Source materials used in the work are presented in individual chapters of the dissertation.

Although the network and traditional library resources contain all sorts of materials from the Big Data field, many problems could not be answered. It seems that there is still a lack of research on the use of new technologies in the Big Data area, including refining information, and in particular identifying and collecting information. Often, due to a lack of literature, new solutions to the problems were sought, which is part of the innovation of the dissertation.

Many sources of information were found on the internet. IT book publications do not always contain up-to-date content, and sometimes include a description of an older version of a program or system. In special cases, to solve a technical problem that prevented a study from being performed, it was necessary to study the software source codes to understand and fully consciously apply the solution in the actual research process. In many places, due to the lack of available literature, reverse engineering was the only way to perform research in the area of social sciences – an area that is not directly related to data processing technology. This shows, among other things, how much information technology is becoming a part of humanities and social sciences. It also further clarified the legitimacy of the work.

[9] A sentiment is understood as a feeling - positive, neutral, or negative.

Construction of doctoral dissertation

The work consists of an introduction, three chapters, a conclusion, annexes, and a bibliography. The introduction to the dissertation defines its purpose, research hypothesis, and assumptions. A literature review of the current state of knowledge of Big Data is given in chapter one. The second chapter describes the Big Data robot. The third chapter presents the subject of the dissertation – the results of four studies based on data collected by the Big Data robot. Finally, conclusions regarding the subject and goal of the dissertation are given.

The first chapter is devoted to issues regarding the potential and needs of communication through a medium such as the network. The history of the internet is outlined, including changes in the way its participants communicate. Methods of communication on the network, including email, USENET servers, IRC, WWW, and social networking sites are discussed. The definition of the concept of Big Data adopted for the purposes of the work is presented. Examples of applications of large information sets are described, taking into account the two main categories of Big Data – structured and unstructured data. The information potential of Big Data is presented, including the potential for it to provide new, useful information from data available on the web. Several examples and benefits of using large data sets are described. The chapter also deals with aspects of data security and maintaining the privacy of network users. This is important as network communication, currently dominated by websites, often includes personal data from participants.

The second chapter discusses the possibility of obtaining information from the network, which is a product of the communication of its users. This communication, regardless of its type (active, direct and synchronous, or direct and asynchronous), leaves large amounts of digital resources on the Internet[10]. These resources can take a variety of storage formats, such as text, sound, image, video, and multimedia as a compilation of the previously mentioned formats.

All digital data is stored in digital information carriers, which can also be a valuable source of information. However, not all data formats are identifiable for the Big Data robot. The Big Data robot is only able to collect texts that are saved to a database. Therefore, other data formats, e.g., content in audio and video formats, must be converted to text. The reason that the robot can only collect textual data is due to the fact that the other stages of data analysis, including refining, involve the processing of textual data. This chapter therefore also discusses methods for converting audio, video, and video content into text. Digital materials that are not available on the network, such as document scans, internal databases, and digital archives, are other interesting sources of data that can be loaded into the robot's database.

[10] W. Gogołek, *Komunikacja sieciowa*...

Content posted on the web, which is a product of interpersonal communication, may be available openly, without restrictions, or hidden from direct reading. In connection with this, this chapter also contains a description of data access safeguards that prevent data from being downloaded by the Big Data robot. It also mentions the rules that should be followed when collecting data and the resources hidden in the Deep Web[11]. The summary of the chapter describes the components of the Big Data robot and how it works.

The third chapter presents the possibilities of conducting research on information obtained from large data sets. The completed research discussed in the chapter was selected in order to show the potential of the network to be used in research and the usability of large amounts of current, historical, structured, and unstructured data available on the internet. This chapter describes four studies that used data collected by the Big Data robot.

The first study described in this chapter is a 2015 survey on presidential and parliamentary elections and data on the same topic collected from the internet by the Big Data robot. In the survey of Poles' electoral preferences, the difference between the votes announced by the National Electoral Commission for the presidential candidates and the results obtained from the refining process differed by 0.66%[12].

The second study was based on the image of John Paul II in the network. In this study, the robot collected historical data, largely from the USENET network, which provides an archive of the content of discussions conducted by network users[13]. After refining the historical data, it was found that Pope John Paul II has a very positive image on the web, and negative entries against him are few[14].

The third study was a non-social study concerning the health of Poles and was carried out using data from health websites. The study confirmed that refining information can be an effective

[11] The Deep Web is hidden internet resources that cannot be found by search engines. Hidden resources include information sources protected by passwords as well as encrypted computer networks.
[12] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Część 1. Blogi, fora, analiza sentymentów*, „Studia Medioznawcze" nr 2(53) (2013).
[13] *USENET | Internet discussion network* [na:] „Encyclopedia Britannica", https://www.britannica.com/technology/USENET, udostępniono 29 września 2017 r.
[14] J. Olędzki i in., *Wielkość czy autorytet?: Jan Paweł II w przekazach polskich mediów podczas jego kanonizacji : praca zbiorowa*, Warszawa 2016, s. 353.

way to support monitoring the health of Poles[15]. In addition, the results obtained were consistent with the WHO report on pathogenic causes.

The fourth study showed the capabilities of the Big Data robot in collecting structured data from one source – a competition page – and detecting voting fraud. In this study, the internet competition 'Support for libraries', organized by Empik in 2016, was analyzed, in which the participants voted for the best school library via Facebook. The study proved that the method of collecting available data and then analyzing them is an effective way of detecting irregular behavior in online voting, despite the fact that the amount of available data was small compared to what the organizer had. Therefore, it is possible to verify the integrity of voting regardless of the organizer.

The studies cited in chapter three illustrate the functionality of the Big Data robot methodology. The findings of the studies show that the Big Data robot is a universal tool for collecting data that can be used as raw material for research.

It should be emphasized that data collection is complex. The Big Data robot should be seen as a system related to the network and its changes. Each of the studies, carried out with the help of this tool, enabled the discovery of previously unknown resources of new knowledge about social phenomena, including politics, science, and entertainment. Each piece of research carried out with the use of this robot brings new experiences, which can help researchers to make the right decisions when planning subsequent analyses and the data collection processes required for them. The Big Data robot allows new areas to be explored.

The end of the dissertation contains conclusions from the conducted research, a reference to the hypothesis, and assumptions for the Big Data robot. It also discusses potential directions of research methodology development in the field of network communication using the robot, as well as possible development possibilities for the robot itself.

[15] S. Młodzianowska, *Rafinacja sieciowa jako narzędzie monitoringu stanu zdrowia Polaków* [w:] Wydział Dziennikarstwa i Nauk Politycznych 2016, s. 74.

Summary

Network communication in the modern world has become the dominant form of information transfer. With the development of civilization, traditional face-to-face conversation is increasingly being replaced by other types of communication, especially network communication: human to human, human to machine, and machine to machine.

Today, humanity is immersed in digital data. Information is created faster than ever before, and the network plays a key role in communicating it. A working data collection system is a fundamental requirement for achieving the goal set out in this dissertation.

The dissertation describes four sample studies in which structured, unstructured, current, and historical data were collected. The conducted research concerned very different areas – Poles' election preferences, the image of John Paul II, causes of diseases in Poland, and fraudulent voting in an online competition. These studies prove the effectiveness of the constructed Big Data robot and the developed method of collecting data from the network.

The studies also show that previously unknown horizons in areas that were up to now inaccessible are open to the Big Data robot. Large data resources, their collection and subsequent refinement, allow new useful information that is hidden from people to be obtained. Structured and unstructured data on any topic can be studied. This is the strength of the Big Data robot. Therefore, it can be used wherever it is required to collect large amounts of data that will allow new information to be obtained during the refining process. The modular design of the robot allows the tool to be adapted to a constantly changing network. This modularity is another attribute of its universality.

The solutions used to build the Big Data robot are largely found in open source software. Thus, there are unlimited possibilities to modify and adapt the robot to current needs and hardware conditions. The robot does not need to be used solely for social research. To a large extent, network data is open information published by public, state, and local government entities, research and development centers, social entities, and companies, and these all provide huge amounts of data to the network. Materials such as reports, projects, patents, and scientific publications can also be a valuable source of information that can be collected and refined. The results of analysis of this type of data can be very useful, and help when making strategic decisions in companies, research centers, or state institutions.

The future of the Jazon Big Data robot can be seen in its compatibility with data generated by other devices connected to the network (IoT), because humans are not the only source of information. The IoT is already a huge resource, which autonomously generates huge amounts of data. A big opportunity for the Big Data robot is artificial intelligence (AI), which is increasingly entering everyday life. The field of science of AI is growing rapidly. The possibility of the self-improvement of robots using AI mechanisms may be the future for many solutions, including the Big Data robot, which could use a neural network to categorize and, in the long run, assess the value of a given source of information.

The functions of the Big Data robot and its ability to process large data sets make it a tool that supports the work of researchers, as it helps with laborious data collection. The advantage of using the Jazon Big Data robot is that it can be used to analyze the entire amount of collected information. Jazon's system has been constructed in such a way that, along with the development of the network and its resources, it can be developed and adapted to new research areas directly related to Big Data.

In summary, information is an essential component of media science. Information is of particular importance to internet resources and breakthrough technologies that collect, exchange, and process information. As a result, a large data resource space is becoming an area of innovative research on obtaining new valuable information.